8

Application of Statistical Methods for Gas Turbine Plant Operation Monitoring

Li Pan
Queen's University Belfast
U.K.

1. Introduction

Within a large modern combine cycle gas turbine (CCGT) power station, it is typical for thousands of process signals to be continually recorded and archived. This data may contain valuable information about plant operations. However, the large volume of data accompanied with inconsistencies within the data often limits the ability to identify useful information about the process. Utilising data mining techniques, such as Principal Component Analysis (PCA) and Partial Least Squares (PLS), it is possible to create a reduced order statistical model representing normal plant conditions. Such a model can then be utilised for fault identification and identifying possible improvements in key performance indicators such as thermal efficiency. Moreover, the gas turbine performance can be affected by changes in ambient conditions. A long term nonlinear PLS techniques can be applied here to investigate the seasonal changes in gas turbine.
In this chapter, an approach to establish a long term statistical model for gas turbine will be given, and the application of the model in fault detection and performance analysis will be demonstrated.

2. The data mining techniques

Within the power station, data are continuously collected and archived representing thousands of data points including temperatures, steam flow rates, pressures, etc. Potentially this data may contain valuable information about unit operation. However, collecting a large amount of data does not always equate to a large amount of information, leading to a lot of databases being regarded as data rich, but information poor. The task of extracting information from data is known as data mining, which is defined as the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data. Data mining is the nontrivial process of extracting valid, previously unknown, comprehensible and useful information from large databases (Weiss and Indurkhya, 1998). Also, data mining is a generic term for a wide range of techniques which include intuitive, easily understood methods such as data visualisation to complex mathematical techniques based around neural networks and fuzzy logic (Wang, 1999; Olaru and Wehenkel, 1999). Applications are found within diverse areas such as marketing (Humby et al., 2003), finance (Blanco et al., 2002) and industrial process control (Martin et al., 1996). However, despite being a widely applied technique, it is reported that three

quarters of all companies who attempt data mining projects fail to produce worthwhile results (Matthews, 1997). Unfortunately, this indicates that the potential of data mining techniques, with regard to the available data is often overestimated than the reality.
The act of data mining is itself part of a larger process known as knowledge discovery in data, KDD, which encompasses not only the analysis of data, but the gathering and preparation of data and the interpretation of results. Extracting knowledge from large data sets can be achieved through exploratory data analysis to discover useful patterns in data, in the form of relationships between variables.
Many techniques are applied as classification tools, to categorise new data following the analysis of a historical data set. In this chapter, the first method discussed in Section 2.1 is machine learning techniques which use a logical induction process to categorise a series of examples, resulting in decision tree and rules set which can be implemented in decision making processes. Typical application areas are fault diagnosis in industria1 machines (Michalski et a1., 1999) and the assessment of power system security (Voumvoulakis, 2010)
Case based reasoning methods as discussed in Section 2.2, are commonly applied to decision making tasks where previous experience is desirable, but may not be available. Case based reasoning provides an inexperienced user with exposure to experiences from others, through a set of historical 'cases', and has been of particular use in areas such as fault diagnosis (Wang et al. , 2008; Yan et al. , 2007) and system design and planning (Hinkle and Toomey, 1995).
Finally, Section 2.3 discusses multivariate statistical techniques, namely principal component analysis and partial least squares regression, which have been successfully applied to a range of applications areas including chemistry (Wold et a1., 1987), manufacturing (Oliveira-Esquerre et a1., 2004), and power system analysis (Prasad et a1., 2007). It is also extended to finance (Blanco et a1, 2002) and medicine (Chan et a1, 2003) area. Principal component analysis and partial least squares regression are particularly popular in the area of chemometrics, where they are employed in the monitoring of processes which generate large and highly correlated data sets (Yoon and MacGregor, 2001; Kourti et al., 1996).

2.1 Machine learning

Machine learning techniques are those that use logical or binary operations to 'learn' a task from a series of examples, such as symptoms of medical or technical problems, leading to the diagnosis of the problem through the use of decision trees and rule sets which classify data using a sequence of 1ogical steps (Michalski et a1., 1999).
Decision trees are simple top down learning structures, which use Boolean classifiers to 'grow' a tree through recursive partitioning of the sample data using the available attributes. The development of a decision tree starts with the inclusion of all the training data in a root node, resulting in both correctly and incorrectly classified data. In order to 'grow' the tree, the data is recursively split by each attribute until all the attributes in the data have been used. Each node in the final tree, known as a leaf, represents a test on one of the attributes, and the branches from the node are labelled with the Boolean outcomes of the test (Quinlan, 1996).
The basic algorithm of building a decision tree, also known as ID3/C4.5 algorithm, follows a down rule (Quinlan, 1993). In the beginning, all the data is collected in the root node, and the data is recursively subdivided into fewer branches by assessing the information gain of

each attribute in the training data to split the data further, until the terminal node which only contains one attribute is obtained (Quinlan, 1993).
Rule induction is achieved using a bottom up structure, starting with a rule that specifies a value for every available attribute on the decision tree, thereby making the rule as specific as possible. This rule is known as the seed and further rules are developed from it by successfully removing attributes one at a time, until more general rules are acquired. Any rule which includes a counter example is regarded as incorrect and is therefore discarded from the process. The rule learning terminates by saving a set of'shortest'rules. Also a new "RBDT-1" algorithm is devolved for learning a decision tree from a set of decision rules that cover the data instances rather than from the data instances themselves. The method's goal is to create on-demand a short and accurate decision tree from a stable or dynamically changing set of rules (Abdelhalim, A. 2009).
The primary advantage of decision trees is that their simplicity makes them very intuitive to users. However, large data sets call result in vast trees which can be 'needlessly' complex resulting in a largely unusable knowledge base：the ideal tree is as small and linear as possible. Due to their simple nature, decision trees are not suitable for more complex data structures and this is demonstrated by trees that, after pruning, still remain too large to be comprehensible.

2.2 Case based reasoning

Case based learning acquires knowledge from solutions to prior problems and employs it to derive solutions to the current problems. Once a current problem occurs, the similar case and previous solution are retrieved and possibly revised to better fit the current problem. The new solution can be retained into the case base in case to solve future problems. As a result, case based reasoning (CBR) systems are effectively used as lookup tables where 'the system' interrogates an indexed database of relevant cases, and one or more similar cases are retrieved and applied to discover an appropriate solution (Watson, 1999).
A significant issue in CBR is indexing, which limits the search space, thereby reducing case retrieval times. There are many methods for indexing, such as check list based indexing, which identifies predictive features for a case (inductive learning methods may be used) and places them on a list which is then used for indexing, and difference based indexing which selects features as indices that best differentiate one case from another. The user can also manually implement an indexing system, and it has been suggested that selection of indices by the user can be more effective than algorithms for practical applications (Kolodner, 1993). The indexed cases can then either be stored sequentially, making the system easy to maintain but slow to query for larger case sets, or using a hierarchical structure, which will organize cases so that only a small subset are considered during retrieval, thereby reducing search times (Smyth et a1., 2001).
CBR is a self-maintaining system, the database of historical events is updated when new cases occurred and adding to the system's problem solving resources. The advantage of CBR is that it does not require a large number of historical data patterns to achieve satisfactory levels of performance：a CBR model may be created from a small number of cases and the case base can be refined over time (Hinkle and Toomey, 1995). CBR is particularly useful when studying data which has complex internal structures when there is little domain knowledge, enabling the sharing of experience.
Despite these benefits, CBR can be unsuitable for large scale applications as retrieval algorithms are inefficient when faced with handling thousands of cases. Maintenance of the

case base, with respect to adding new cases and the removal of out date cases, may also be a problem as it is largely left to human intervention (Watson and Marir, 1994).

2.3 Multivariate statistical techniques

Statistical methods are employed to analyse the relationships between individual points in a data set, determining characteristics such as the average value and distribution of the data. The simple statistical measure represents a univariate approach to data analysis, which lacks the ability to constructively analyse large, multivariate data sets as the interactions between variables are ignored (Martin et a1., 1996). In contrast, multivariate statistical analysis describes methods capable of observation and analysis of the multiple variables required for system monitoring (Kourti and MacGregor, 1995). This section discusses the multivariate techniques, principal component analysis (PCA), least squares regression and partial least squares (PLS), as they are more suitable to the analysis of large data sets than univariate methods.

Principal component analysis (PCA) is a statistical technique useful for identifying underlying systematic structures in data and separating it from noise (Wold et al., 1987). The identification of patterns in data structures allows PCA to be applied to problems requiring a reduction in the dimensionality of a data set, for example image processing(Bharati et a1., 2003), or monitoring of industrial processes including chemical and microelectronics manufacturing processes (Wise and Ricker, 1991；MacGregor and Koutodia, 1995). These objectives are achieved by transforming variables, which are assumed to be correlated, into a smaller number of uncorrelated variables called principal components (PCs), providing a simpler description of the data structure. Each successive PC accounts for the most significant variability in the data in a particular direction, with the reduction in dimensionality achieved, the original data set can be represented by few PCs.

PCA is a useful tool when large data sets containing highly correlated variables are to be managed. PCA achieves reductions in data dimensionality, thereby simplifying future observation of variables：plotting a few PCs is significantly more convenient than plotting all original variables. Furthermore, the comparison capabilities between the historical information used to construct the model and newly presented data is a desirable characteristic for system monitoring applications (Martin et a1., 1996). Fault identification can also be undertaken by analyzing the contribution of the independent variables to each PC (MacGregor et a1., 1994).

Projection to latent structures, also known as partial least squares (PLS), is developed to solve the multi-collinearity problem in linear least squares regression (LSR) which can determine the best linear approximation for a set of data points (Freund and Willson, 1997).

The benefit of PLS is achieved by identifying a set of uncorrelated, latent variables. This avoids the co-1inearity problems encountered by likelihood-ratio (LLR) test and utilises some of the techniques associated with PCA, with the new latent vectors composed of scores and orthogonal loading vectors. PLS regression is a robust, multivariate linear regression technique which is considered to be more suitable for the analysis and modelling of noisy and highly correlated data than LLR as parameters do not exhibit large variation when new data samples are included (Otto and Wegscheider, 1985). A high number of variables, with respect to the number of data samples, are also permissible in PLS, which can result in the modelling of noise for LLR(wise and Gallagher, 1996).

In summary, PLS is capable of producing robust, effective models, despite operational data limitations, for example, imprecise measurements and missing data (Oliveira-Esquerre,

2004). The ability to predict dependent data values, especially in the case of product quality data, which is often measured infrequently, is useful in process monitoring (MacGregor, 2005). The proficiency of PLS dealing with the highly correlated and collinear data is also frequently utilized in its application to process monitoring (Kresta, 1994).

2.4 Method selection

A selection of data mining techniques has been presented in this section and the characteristics of each shall now be considered with respect to the problems and available data presented by power plant monitoring and analysis.

Data derived from power plant monitoring can potentially consist of thousands of sensor measurements generated on a second by second basis. The data recorded is highly correlated, due to multiple sensors, which are in place to introduce redundancy into the measurements, and parallel paths within the system, for example, the steam and gas circuits. Data quality is also a factor, with noisy signals and missing data as the common problems. The correlated structure and data quality considerations present in power plant monitoring records bear strong similarities to other application areas discussed in this section, such as chemical process control(Ma et al., 2009; Ahvenlampi and Kortela, 2005), manufacturing (Oliveira-Esquerre et al., 2004; Baharati et a1., 2004; Hisham et al., 2008) and medicine (Chan et a1., 2003).

When monitoring a process which records a vast array of sensor data, individual analysis of each signal by a human operator is clearly not possible. The data analysis techniques discussed in this section can be applied to this problem as they identify essential correlations within the data. This simplifies the monitoring process by identifying the most relevant process signals and thereby reducing the search space.

When undertaking system monitoring, there are three main objectives. The first is the detection of a change in process operation and its nature - should this be a sensor fault, faults within the process or a change in product quality or system performance. Once a change in process behaviour is detected, diagnostic tests are then required to identify the cause of the change, which may require analysis of recent process data and/or consultation with an experienced operator. Finally, with the source of the change ascertained, a solution should then be identified and implementation of corrective action undertaken, if appropriate (Cinar and Undey, 1999). This may require either disabling the source of the problem, or in the event of a faulty sensor reading, reconstruction of the value.

The methods presented here provide different solutions to the process monitoring problem. Clustering, machine learning and CBR are diagnostic tools, which compare current fault conditions to historical examples of faults. While these techniques will identify the nature of the problem, they are not capable of detecting its occurrence or providing signal reconstruction in the event of a sensor fault. Statistical methods offer a model based approach, where process operation data is compared to a system model, based on 'normal' operating conditions. This provides continuous process monitoring, which can supply early warning of a small process change and offers operators the opportunity to take action to prevent the fault becoming more serious. In particular, PCA and PLS can supply the operator with information as to which process variables are outside normal limits (MacGregor et a1., 1994). This serves to focus the operators' attention on the problem area, allowing process knowledge to be used to identify the source of the problem.

Not all of the techniques discussed in this section are capable of providing a solution that enables preventative action to be taken for CCGT power plant.

CBR, providing a historical event that is similar to the currently observed event, is also of little use in this instance, as it again offers classification of new events, however, an explanation of the similarities identified between cases is unavailable.
Cluster analysis and 'rules' may be useful in identifying groups of similar events, as their aim is to suggest correlations in data. Once similarities have been identified for events resulting in groupings with common characteristics further investigation would be required to identify the relationships between variables that cause these groupings. A more complete solution is offered by statistical methods.
Both PCA and PLS are capable of identifying correlations within data, while PLS also offers the ability to extend this to identifying the correlations which are predictive of a dependent quantity. The correlations identified within the data can then be studied using the scores and loading vectors obtained, indicating the contribution of variables, if any, to the variation of the dependent parameter. The historical data is suitable for the development of a system model, by means of PCA and PLS, which can be applied to continuous monitoring.
Archived data is available, detailing sensor data, for example temperatures, pressures, etc, throughout the plant at regular intervals. Once the PCA and PLS models are developed, it can provide a relatively straight forward model which has both the ability for online fault monitoring and offline performance analysis. For practical application, those PCA and PLS models are required for fast online response within 20 seconds, and reasonable prediction accuracy in a wide operation range.
With PCA and PLS identified as possessing properties that are useful in relation to the problems posed by power plant and power system operation, the statistical modelling methods will provide the most suitable approach for operation monitoring and performance analysis of CCGT power station.

3. PCA and PLS algorithm

Give an original data matrix $\mathbf{X}$ $(m \times n)$ formed from m samples of n sensors, and subsequently normalised to zero mean and unit variance, can be decomposed as follows:

$$\mathbf{X} = \mathbf{T}\mathbf{P}^T + \mathbf{E} = t_1 p_1^T + t_2 p_2^T + ... + \mathbf{E} = \mathbf{PC}_1 + \mathbf{PC}_2 + ... + \mathbf{E} \quad (1)$$

where $\mathbf{T} \in R^{m \times A}$ and $\mathbf{P} \in R^{n \times A}$ are the principal component score and loading matrices, E is the residual matrix (Lewin, 1995).
The principal component matrices can be obtained by calculating eigenvectors of original data. Following the creation of the correlation matrix of original data, the corresponding eigenvalues and eigenvectors are calculated, where an eigenvalue is an eigenvector's scaling factor. As the eigenvectors with the largest eigenvalues correspond to the dimensions than have the strongest correlation in the dataset, the data can then be ordered by eigenvalue, highest to lowest to give the components in order of significance (Jolliffe, 2002). There are a number of methods available to determine the number of ordered PCs. A cross validation which calculates the predicted error sum of squares (PRESS) (Valle et a., 1999) is provide more reliable solutions than a simple scree test (Jackson, 1993).
Partial least square requires two block of data, an **X** block (input variables) and **Y** block (dependent variables). PLS attempts to provide an estimate of **Y** using the **X** data, in a similar manner to principal components analysis (PCA). If **T** and **U** represent the score matrixes for the **X** and **Y** blocks, and **P** and **Q** are the respective loadings, the decomposition equations can be presented as:-

$$\mathbf{X} = \mathbf{T}\mathbf{P}^T + \mathbf{E} \tag{2}$$

$$\mathbf{Y} = \mathbf{U}\mathbf{Q}^T + \mathbf{F} \tag{3}$$

where **E** and **F** are the residual matrices. If the relationship between **X** and **Y** is assumed to be linear then the residual matrices **E** and **F** will be sufficiently small, and the score matrices **T** and **U** can be linked by a diagonal matrix **B** such that:

$$\mathbf{U} = \mathbf{B}\mathbf{T} \tag{4}$$

Hence the predicted dependent variable can be translated (Flynn, 2003) as:

$$\hat{\mathbf{Y}} = \mathbf{B}\mathbf{T}\mathbf{Q}^T + \mathbf{E} + \mathbf{F} \tag{5}$$

4. Nonlinear modeling approach

As we discussed previously, PCA and PLS model are powerful linear regression techniques. However, in the real power generation industry, many processes are inherently nonlinear. When applying linear model to a nonlinear problem, the minor latent variables cannot always be discarded, since they may not only describe noise or negligible variance structures in the data, but may actually contain significant information about the nonlinearities. This indicates that the linear model may require too many components to be practicable for monitoring or analyzing the system.

Recognition of the nonlinearities can be achieved using intuitive methods, for example, which apply nonlinear transformations to the original variables or create an array of linear models spanning the whole operating range. More advanced methods have also been proposed including nonlinear extensions to PCA (Li et al. 2000), introducing nonlinear modifications to the relationship between the **X** and **Y** blocks in PLS (Baffi et al., 1999) or applying neural network, fuzzy logic, etc. methods to represent the nonlinear directly.

Transformation of the original variables using nonlinear functions can be introduced prior to a linear PCA and PLS model. For this purpose, the input matrix **X** is extended by including nonlinear combinations of the original variables. However, process knowledge and experience is required to intelligently select suitable nonlinear transformations, and those transforming functions must sufficiently reflect the underlying nonlinear relationships within the power plant. Another problem with this approach is the assumption that the original sets of variables are themselves independent. This is rarely true in practice, which can make the resulting output from the data mining exercise difficult to interpret.

An alternative and more structured approach is the kernel algorithm. The purpose of kernel algorithm is to transform the nonlinear input data set into a subspace with kernel function. In the kernel subspace, the nonlinear relationship between input variables can be transformed into linear relationship approximately. By optimising the coefficients of kernel function, the transformed data can be represented using a Gaussian distribution around linear fitting curve in the subspace. Furthermore, introducing neural network approaches into the kernel structure is generally seen to be more capable of providing an accurate representation of the relationship for each component (Sebzalli and Wang, 2001). In this area, the multilayer perceptron (MLP) networks are popular for many applications. However the initial model training is a nonlinear optimization problem, requiring conjugate

gradient and Hessian-based methods to avoid difficulties arising from convergence on local minima. In order to solve this problem, a radial basis function (RBF) network has been selected over other approaches, due to its capability of universal approximation, strong power for input and output translating and better clustering function. A standard RBF network consists of a single-layer feedforward architecture, with the neurons in the hidden layer generating a set of basis functions which are then combined by a linear output neuron. Each basis function is centered at some point in the input space and its output is a function of the distance of the inputs to the centre. The function width should be selected carefully because each neuron should be viewed to approximate a small region of the input surface neighboring its centre. Therefore, the RBF network also has been named localized receptive field network. This localized receptive character implies a concept of distance, e.g. the RBF function is only activated when the input has closed to the RBF network receptive field. For this reason, the performance of RBF network is more dependent on the optimisation of RBF function coefficients rather than the type of function (Jiang et al., 2007).
In order to reduce the neural network dimension, the input data are firstly decomposed into few components, then the output can be reconstructed with nonlinear relationship. Hence, each component will possess its own nonlinear function $f_{non-linear}$, so that

$$\hat{\mathbf{X}} = f^{\mathrm{X}}_{non-linear}(\mathbf{T}) \tag{6}$$

$$\hat{\mathbf{Y}} = f^{\mathrm{Y}}_{non-linear}(\mathbf{T}) \tag{7}$$

In this research, radial basis functions have been selected to represent the non-linearities, since once the RBF centres and widths have been chosen, as described below, the remaining weights can be obtained using linear methods.

4.1 RBF network

The radial basis function network employed in this research is illustrated in Figure 1.

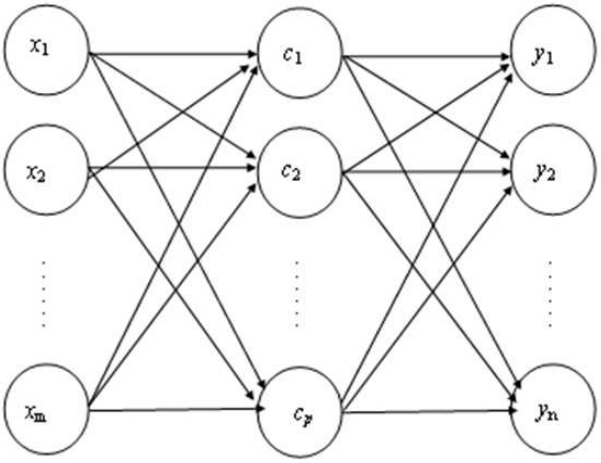


Fig. 1. Radial basis function network

The network topology consists of *m* inputs, *p* hidden nodes and *n* outputs, and the network output, y_i , can be formulated as:-

$$y_i = \sum_{j=1}^{p} w_j^{(i)} \theta_j \left(\left\| \mathbf{X} - \mathbf{c}_j \right\| \right) \quad i = 1, 2, \ldots n \tag{8}$$

where, $w_j^{(i)}$ are weighting coefficients, and θ_j is the basis function. In this research, a Gaussian base function was selected, which is defined as:-

$$\theta_j \left(\left\| \mathbf{X} - \mathbf{c}_j \right\| \right) = \exp\left[-\sum_{k=1}^{m} \left(\frac{x_k - \mathbf{c}_j}{\sigma_j} \right)^2 \right], \; i = 1, 2, \ldots p \tag{9}$$

The Euclidean distance $\left\| \mathbf{X} - \mathbf{c}_j \right\|$ represents the distance between the input space **X** and each RBF centre $\mathbf{c}_j$, where **X** = [$\mathbf{x}_1$ $\mathbf{x}_2$ … $\mathbf{x}_m$], and σ_j is the width coefficient of each RBF node. The coefficient matrix [σ , c, w] is obtained off-line using a suitable training algorithm. Some of the more popular options are least mean squares (LMS) (Moody et al., 1989), orthogonal least squares (OLS) (Li et al,. 2006) and dual-OLS (Billing et al., 1998). These traditional algorithms often employ a gradient descent method, which tends to converge on local minima. In order to address the global optimisation problem, a recursive hybrid genetic algorithm (RHGA) (Li and Liu, 2002, Pan et al., 2007) is employed here to search for valid solutions.

4.2 The genetic algorithm

The typical genetic algorithm (GA) is based upon *survival of the fittest*, and the network framework [σ , c] is coded into the binary genes as illustrated in Table 1. The initial population are selected at random from the entire solution space, with the binary coding denoting whether the training samples are selected as the centers of the hidden neurons (Goldberg, 1989).

All the potential hidden centers	A randomly created gene code	Coded network framework
[σ_1, $\hat{c}_1$]	1	[σ_1, $\hat{c}_1$]
[σ_2, $\hat{c}_2$]	0	---
[σ_3, $\hat{c}_3$]	0	---
[σ_4, $\hat{c}_4$]	1	[σ_4, $\hat{c}_4$]
[σ_5, $\hat{c}_5$]	1	[σ_5, $\hat{c}_5$]

Table 1. Encoding scheme of genes

For each generation, random crossover and mutation is applied to the genes, leading to a new generation of network frameworks being obtained. The fitness, *f*, of the new population is determined using:-

$$\frac{1}{f} = \sum_{j=1}^{n} (\hat{y}_j - y_j)^2 \tag{10}$$

where, $\hat{y}_j$ is the j^{th} RBF output and y_j is the actual value. The most recent framework will be retained if its fitness improves upon previous generations.

Although the genetic algorithm has the capability of wide region searching and efficient global optimizing, it is weak in some local point fitting. This may lead to a decrease in model accuracy. Therefore, the genetic and gradient descent algorithm can be combined in order to obtain both the global and localize optimizing capability (Pan, et al., 2007). In this hybrid algorithm, an initial optimized network can be obtained by the genetic algorithm, and then the structure of network can be further shaped for some specific points with the gradient descent algorithm. The next step is to examine the variate of fitness coefficient. If the fitness reached the preset bound then the regression will be completed, otherwise, the network will be reconstructed for next generation optimisation, and repeat the gradient descent regression, until reach the preset number of generations or meet the request fitness.

5. The auxiliary methods

Once a PCA/PLS model for normal operating conditions has been developed, the real time online DCS data then can be applied into the model to obtain a reconstruction of input data. It can be used to determine whether recorded plant measurements are consistent with historical values and neighboring sensors. A comparison can then be made between the reconstructed value for each variable and the actual measurements. Performed manually this can be a time consuming task. In this section, some efficient auxiliary methods will be discussed for the quality control, sample distribution analysis and fault identification.

5.1 Quality control method

There are two approaches that can quickly help to identify differences between the actual and reconstructed value of a variable, which are the squared prediction error (SPE) and Hotelling's T^2 test.

The SPE value, also know as the distance to the model, is obtained by calculating a reconstruction of each variable, $\hat{x}_i$, from the model, and then comparing it with the actual value, x_i. The SPE for all variables in each data sample can be calculated as

$$\mathbf{SPE} = \sum_{i=1}^{n} (x_i - \hat{x}_i)^2 \tag{11}$$

In order to distinguish between normal and high values of SPE, a confidence limit, known as the Q statistic test is available, which can be determined for α percentile confidence as:

$$\mathbf{Q}_\alpha = \theta_1 \left(\frac{c_\alpha \sqrt{2\theta_2 h_0^2}}{\theta_1} + 1 + \frac{\theta_2 h_0 (h_0 - 1)}{\theta_1^2} \right)^{1/h0} \tag{12}$$

where c_α is the confidence coefficient for the 1– α percentile of a Gaussian distribution, θ_i is the sum of unused eigenvalues to the i^{th} power and h_0 is a combination of θ as outlined below:

$$h_0 = 1 - \frac{2\theta_1 \theta_3}{3\theta_2^2} \tag{13}$$

The T^2 statistic test is designed as a multivariate counterpart to the student's t statistic. This test is a measure of the variation within normal operating conditions. With Tracy- Widom distribution, the T^2 test can be extended to detect peculiar points in the PCA model (Tracy et al., 1993).
Given h components in use, t_i is the i^{th} component score and s_i is its covariance, then the T^2 can be defined as

$$\mathbf{T}^2 = \sum_{i=1}^{h} \frac{t_i^2}{s_i^2} \tag{14}$$

As with SPE, an upper control limit, T_α^2 can be calculated with n training data. This relates the degrees of freedom in the model to the F distribution,

$$\mathbf{T}_\alpha^2 = \frac{h(n^2-1)}{n^2(n-h)} \mathbf{F}_{1-\alpha}(h, n-h) \tag{15}$$

It should be noted that a rise in the SPE or T^2 value does not always indicate a fault, it also may be caused by the process is moving to a new event which is not accounted in the training data. Additionally, both indicators are affected by noise on the system and deviation of measurements from a normal distribution. This can result in nuisance values for both SPE and T^2. However, false alarms can be largely eliminated by simple filtering, and adjustment of the associated threshold (Qin et al., 1997).

5.2 Sample distribution

Both the SPE and T^2 are unlikely to differentiate between a failing sensor and a fault on the power plant. In this case, a plotting of t scores can be combined with the previous methods to distinguish between the two conditions.
The PCA model gives a reduction of data dimension with minimum information less. Therefore, the original m dimension data can be plotted in a plane coordinated by the first two components, and the relative position between each data point is remained the same as the original m dimension space. This character gives a capability to directly observe the similar distribution structure of original sample data, in a 2-dimension plane.
Especially, quoting the T^2 control limit into the 2-dimension plane, we have

$$\mathbf{T}_{2-D}^2 \le \mathbf{T}_{\alpha\,2-D}^2 \tag{16}$$

substituting Eq. (14) and (15), the Eq. (16) can be transformed as

$$\left(\frac{t_1^2}{s_1^2} + \frac{t_2^2}{s_2^2} \right) \le \frac{2(n^2-1)}{n^2(n-2)} \mathbf{F}_{1-\alpha}(2, n-2) \tag{17}$$

Define

$$c = \frac{2(n^2-1)}{n^2(n-2)} \mathbf{F}_{1-\alpha}(2, n-2) \tag{18}$$

then it gives that

$$\frac{t_1^2}{s_1^2} + \frac{t_2^2}{s_2^2} \le c \tag{19}$$

Eq. (19) defines a control ellipse for *t*-score plotting. Score for normal operating conditions should fall within this ellipse. So when a process fault occurs, the individual points on the *t* score plots may be observed drifting away from the normal range into a separate cluster. The relative position of these fault clusters can assist in latter diagnosis.

5.3 Fault orientation

Having confirmed that there is a sensor fault, and not a process condition, the next step is to identify which sensor is failing. If a signal is faulty, a significant reduction in SPE before and after reconstruction would be expected. However, in practice the reduction in SPE can affect all inputs, making the faulty sensor unidentifiable. This situation arises due to a lack of redundancy, or degrees of freedom, among the measurements.

The above difficulties can be overcome by calculating a sensor validity index (SVI) (Dunia et al, 1996). This indicator is determining the contribution of each variable to the SPE value. The SPE value should be significantly reduced by using the reconstruction to replace the faulty input variable. If an adjusted data set z_{i} represents a input set with the x_{i} variable being replaced by reconstructed data $\hat{x}_i$, and the adjusted model predicted value being $\hat{z}_i$, then the sensor validity index for i^{th} sensor η_i can be defined as

$$\eta_i^2 = \frac{(z_i - \hat{z}_i)^2}{\mathbf{SPE}} \tag{20}$$

The SVI is determined for each variable, with a value between 0 and 1 regardless of the number of samples, variables, etc. The value of SVI close to unity is indicative of a normal signal, while a value approaching zero signifies a fault. It is assumed that a single sensor has failed, and the remaining signals are used for reconstruction. Also, system transients and measurement noise can lead to oscillations in SVI, and possibility of false triggering. Consequently, each signal should be filtered and compared with a user-defined threshold.

6. Application of PCA and PLS model

As these power plants operate in a competitive market place, achieving optimum plant performance is essential. The first task in improving plant operation is the enhancement of power plant operating range. This power plant availability is a function of the frequency of system faults and the associated downtime required for their repair (Lindsley, 2000). As such, availability can be improved through monitoring of the system, enabling early detection of faults. This therefore allows the system working at non-rated conditions, corrective actions, or efficient scheduling of system downtime for maintenance (Armor, 2003).

Monitoring of power plant operations is clearly an important task both in terms of identifying equipment faults, pipe leaks, etc. within the generating units or confirming sensor failures, control saturation, etc. At a higher level, issues surrounding thermal efficiency and emissions production for each generating unit, as measures of plant performance, and the seasonal influence of ambient conditions will also be of interest. Fortunately, the frequency of measurement and distribution of sensors throughout a power

station provides a great deal of redundancy which can be exploited for both fault identification and performance monitoring (Flynn et al., 2006). However, modern distributed control systems (DCSs) have the ability to monitor tens of thousands of process signals in real time, such that the volume of data collected can often obscure any information or patterns hidden within.

Physical or empirical mathematical models can be developed to describe the properties of individual processes. However, there is an assumption that faults are known and have been incorporated into the model. This can be a time-consuming exercise and requires the designer to have extensive knowledge of the application in question (Yoon and MacGregor, 2000). Alternatively, data mining is a generic term for a wide variety of techniques which aim to identify novel, potentially useful and ultimately understandable patterns in data. The most successful applications have been in the fields of scientific research and industrial process monitoring, e.g. chemical engineering and chemometrics (Ruiz-Jimenez et al., 2004), industrial process control (Sebzalli et al., 2000) and power system applications such as fault protection in transmission networks (Vazquez-Martinez, 2003). In the following sections it will be shown how using the principal component analysis (PCA) technique. It is possible to exploit data redundancy for fault detection and signal replacement, as applied to monitoring of a combined cycle gas turbine.

Furthermore, the archived data is used to assess system performance with respect to emissions and thermal efficiency using a partial least square (PLS) technique.

6.1 Raw data pre-process

The PCA and PLS models are trained using historical data to suit the 'normal' plant operating, and the training data have to be selected carefully to avoid failing and over range data from normal power plant operation. The normal power plant operation was defined around the typical output range of 60 MW – 106 MW for single shaft unit and 300 MW – 500 MW for multi-shaft unit. There are severe dynamic conditions existing in the starting up and shutting down period. Therefore, those periods has to be removed from raw data archives. An instance is illustrated in Figure 2, for a single shaft unit operation, approximately one hour operating data was removed after and before system shut down and start up, in order to avoid the transient process.

The DCS normally collects sensor data every second, however, due to the power plant parameters are mainly consisted by temperature and pressure signals, the typical power plant responding time is around minutes. Therefore, consider of the balance of computational complexity and information quality, the sampling interval was determined as 1 minute. Since the raw data sample was archived from DCS, it still contains lots of anomalous signals such as break down process, which the power out suddenly crash down. Noised signal, is a signal disturbed by white noise. And spike, is an instantaneous disturbance which can cause a far deviation from normal signal level. Those data must be pre-filtered before being employed to train a model.

It is generally recognized that CCGT performance, and in particular gas turbine performance, can be affected by changes in ambient conditions (Lalor and O'Malley, 2003). For example, a fall in barometric pressure causes a reduction in air density and hence inlet compressor air flow. Similarly, an increase in ambient temperature causes a reduction in air density and inlet compressor air flow. Since the turbine inlet temperature is maintained as a constant, there is a subsequent reduction in turbine inlet pressure and hence cycle efficiency.

Variations in other external variables such as relative air humidity and system frequency (affecting compressor rotational speed) can also impact on gas turbine performance. Therefore, the training data selection for a widely suitable PCA model has to contain the information of the seasonally changes of ambient condition.

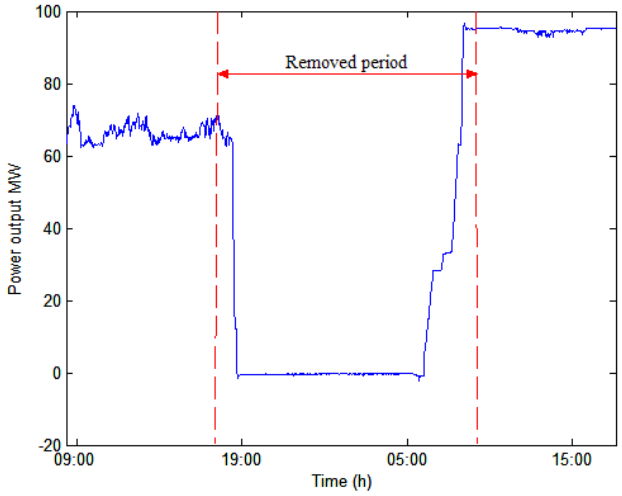


Fig. 2. Removed transient period

In order to obtain a entire seasonal model, the training data sorting process is designed to archive power plant operating data for years, then split all of the ambient variables into many small intervals, and pick up a sample data from each intervals to ensure that the training data contain the operating information for every ambient conditions.

6.2 Sensor data validation

With aging sensors, and the associated performance degradation, inevitable, faulty sensors are a relatively common occurrence in system monitoring. A common example of sensor failure is 'stuck at' signal, as illustrated in Figure 3 (a), which the fault is occurred at 300th data point. The following data is missed and the sensor's output is stuck at the last measurement. Another example is drifting signal, shown as Figure 3 (b), that the original data is disturbed by an increasing interference. Also, a biased signal is a constant noise which biased the sensor's data to other level, as shown in Figure 3 (c).

Univariate limits, i.e. upper and lower bounds are often applied to the detection of these faults. Problems such as biased sensors can be detected when the value eventually exceeds the predefined limits. However, a faulty signal within the univariate limits, such as a drifting sensor, will often go undetected for a long period of time. In order to identify such those faulty sensors, a multivariate approach is required, which will give consideration to the sensor value as part of wider plant operation.

Furthermore, if a sensor is faulty, an operator may choose to disable the sensor, but if the signal is used for feedback/feedforward control, disabling the sensor can only be part of the solution. In this instance, the problem can normally be resolved by signal reconstruction based upon sensor readings from neighboring sensors in the plant. This will require a system model, operating in parallel with the real plant.

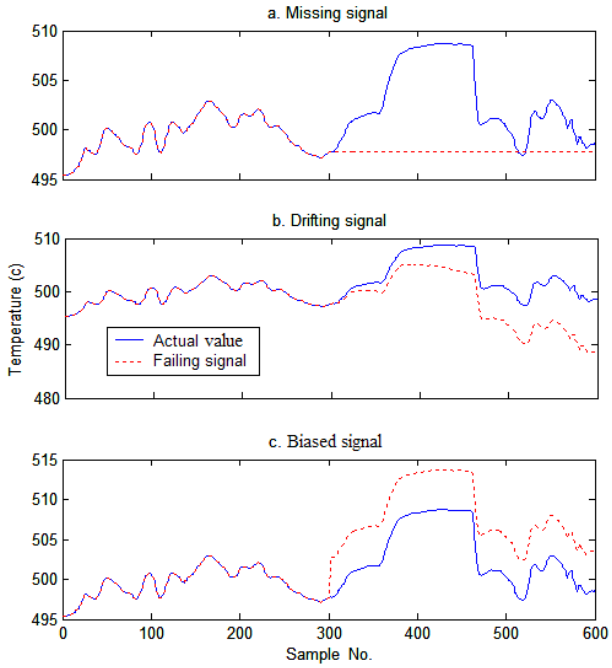


Fig. 3. Sensor faults

Principal component analysis, PCA, as a suitable technique for sensor monitoring and validation as it captures data variability for normal process operation. The development of a PCA model is intended to reduce the dimensionality of a set of related variables, while retaining as much of their variance as possible. This is achieved by identifying new, latent variables known as principal components, PCs, which are linearly independent. A reduced set of these latent variables are then used for process monitoring, with a small number of components normally sufficient to capture the majority of variability within the data.

Monitoring of a system using PCA is a modeling based approach, achieved by comparing observed power plant operation to that simulated by the model from available sensor data. The comparison between model and plant data, resulting in residuals, can then determine if the recorded information is consistent with historical operation and neighboring sensors. Faults are detected by observing deviations from normal operation, which can then be investigated to determine the exact source of the problem.

There are two common automated methods to compare recorded data with the model, as defined in section 5.1, the squared prediction error, SPE, and Hotelling's T2 test. Also, the sensor validity index, SVI, will identify failing sensors, and t score plots, from a cluster representing normal, fault free operation. All of those techniques are detailed in section 5. If an individual sensor is identified as being at fault, it can be replaced with a value reconstructed by the PCA model from other sensor data. However, if the fault is actually with the power plant, corrective maintenance or other necessary action should then be scheduled.

6.3 PCA model performance

In order to demonstrate the monitoring capabilities of this PCA model, a drift signal is introduced to the testing data set. As shown in Figure 4, the drift occurred in the sensor monitoring the steam temperature at 5:00 am. Generally, the lower bound of steam temperature is 500 ºC during power plant normal operating period. Consequently, this drift can be detected by under limit indicator approximately 2 hours after the drift was introduced. In contrast to sensors limit indicator, the associated squared prediction error (SPE) monitoring test is illustrated in Figure 5 and shows that the SPE test detects the sensor fault 30 minutes after the introduction of the drift with 95% confidence limit, and 45 minutes with 99% confidence limit. Similarly, the T-squared test detected this sensor fails within 35 minutes using 95% confidence limit, and it crossed the 99% threshold 10 minutes later, as shown in Figure 6. The earlier SPE and T-square fault identification can provide more time for the power plant operator to take actions to solve problems.

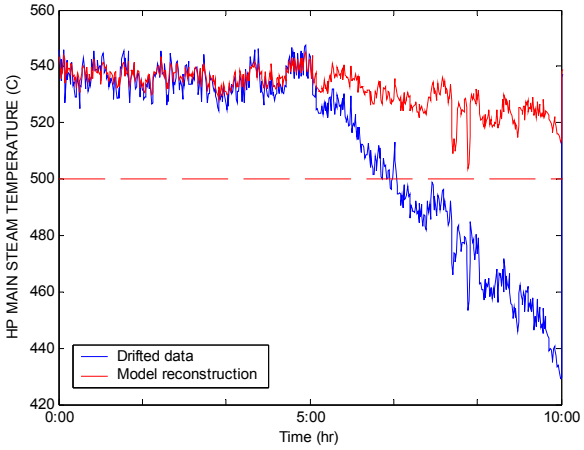


Fig. 4. Sensor drifts for single shaft CCGT unit

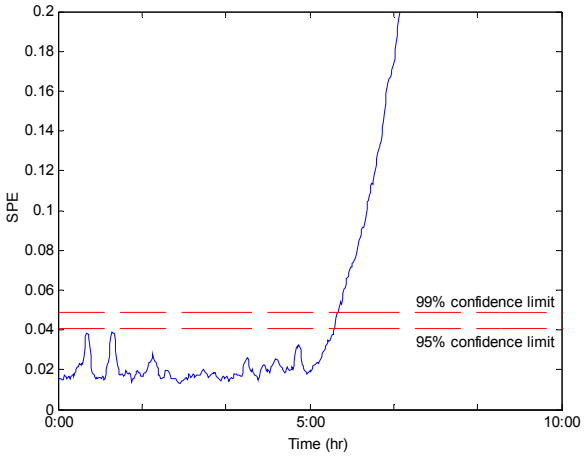


Fig. 5. SPE test for sensor drift in single shaft CCGT unit

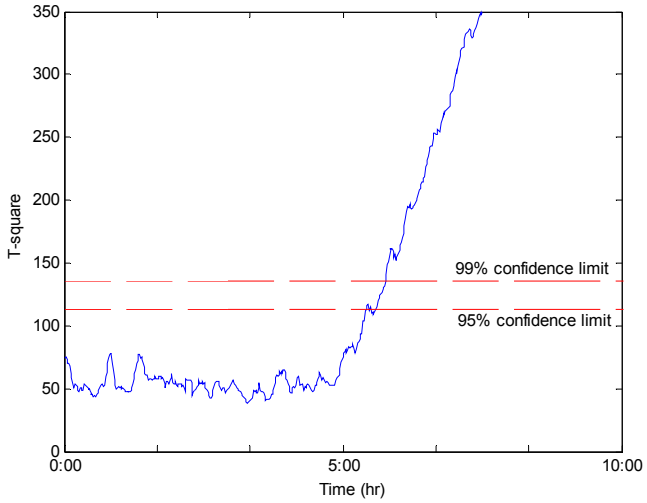


Fig. 6. T-square test for sensor drift in single shaft CCGT unit

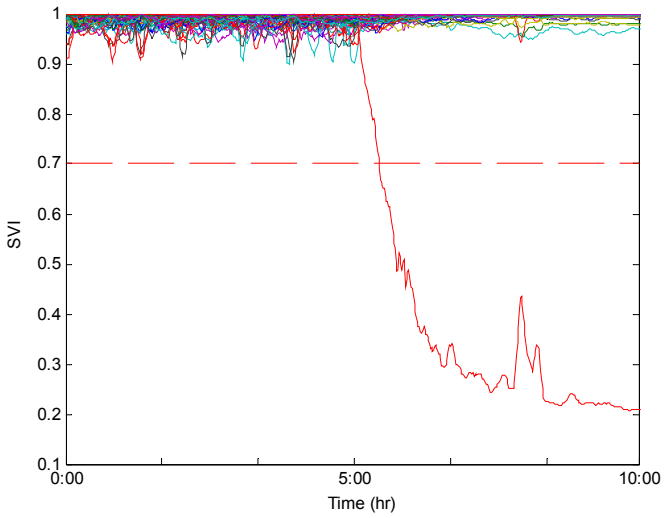


Fig. 7. SVI for sensor drift in single shaft CCGT unit

Following the detection of sensor fault condition, the source of the problem must be identified. Calculation of the sensor validity index (SVI), described in section 5.3, the variations in the SVI for each sensor are illustrated in Figure 7. According to the defined threshold of 0.7, the SVI chart clear identified the faulted sensor at 5:40 am, with the associated index of this signal falling into the range 0.7 to 0.2. Also, system transients and measurement noise can lead to oscillations into the SVI and there is a clearly example of SVI

oscillations caused by system transient during 8:00 am to 9:00 am. It should be noted that when the HP main steam temperature signal drifts, the associated indices for the remaining sensors rise toward unity, accentuating identification of the biased sensor. As the fault is with the sensor, not the process, the PCA model can undertake reconstruction of the failed sensor as shown in Figure 4.

6.4 PLS model performance

Having validated the observed sensor data with PCA processor, optimisation of power plant performance should now be addressed. In order to maximize the power generation and simultaneously minimize the fuel consumed and pollution emissions, the performance variables must be able to monitor online and the internal relationship between the performance variables and associated operating parameters should be able to examine through offline analysis.

However, recover some performance variables is a multi-dimensional problem, such as the thermal efficiency, which depends on power demand, supplied fuel type, even the ambient conditions. Due to the expense and complexity of performance variables monitoring, development of online performance monitoring, capable of determine power plant performance from a variety of process variables, is often desirable. Validated and archived plant data can be employed to develop models which are capable of predicting the quality of process operation while providing an insight into the relationship between quality and associated process conditions.

PLS as a suitable technique for plant monitoring and shall be implemented here to demonstrate how system data can be applied to obtain a model of normal plant operation, with respect to a variety of quality variable measures, such as power plant efficiency, emissions and so on.

As with PCA, monitoring of individual fault conditions is not necessary and problems are instead detected as deviations from normal operation. With load cycling of generation plant increasingly common, a wide range of operating conditions are detailed in archived plant data and potentially contain indicators of operating conditions which lead to optimal power plant performance. The availability of operator logs makes it possible to indentify period of generation regarded by operators to be representative of fault-free power plant performance.

6.4.1 Variance explanation contribution

A benefit from the PLS model is that it has the ability to examine the effect of each input variable on the quality variables. Since the PLS model determines the variance explanation contribution of each variable by examining the correlation to the output variables, the PLS model is not only able to find those variables which have the greatest effect on output, but also can find the variables have indirect effect on the quality variables. This function can be applied to research the effect of any variable we interested, such as air temperature, sea water temperature, humidity and so on.

For instance, the variable contributions to the variance explanation of efficiency are charted in Figure 8 for a normal CCGT plant. Since the input variables are selected for highly related to the efficiency, most of them have comparatively high value of variance explanation, and these can be considered to be important variables to be monitored and/or adjusted when attempting to achieve enhanced operating goals. The most important variables are varying

similar in both single shaft and multi shaft unit. For example, it can be observed that the No.1 and 2 are outstanding with 87.2% and 86.8% variance explained in the single shaft model, and they pointed to the signals of power output and gas flue flow respectively. Contrasts to multi-shaft model, above variables are identified as parameter No.1 with 85.0% explanation for power output and No.6 with 83% explanation for gas flue flow. Also, a group of sensors measuring the high pressure steam parameters are significant, which is the No.19-21 in the single shaft model with around 85% contributions and No. 27-29, 60-62 for both gas turbines in the multi-shaft model with around 80% contributions.

In addition, variations in ambient conditions is also interested, the last 4 variables in both models represent the effects of humidity, air temperature, barometric pressure and sea water temperature, respectively. It is significant that the sea water temperature has an extremely high effect on the power plant efficiency. The reason is considered of the condensing with sea water. The cooler sea water increases heat transfer from the condensing steam, and hence increase the thermal efficiency.

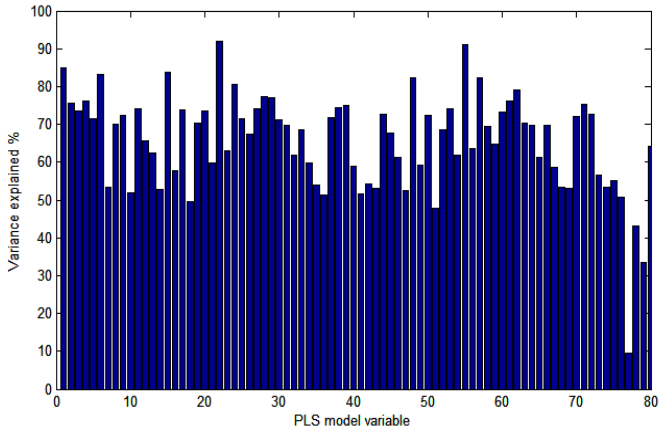


Fig. 8. Variance explanation for CCGT efficiency in multi-shaft unit

6.4.2 Relationship curve

From previous section, the PLS variance explanation suggest that sea water temperature is the most significant ambient condition for thermal efficiency. In order to better appreciate the impact of these environmental variables on the model, we introduce a new technique to study the relationship between input and output variables. It is of interest to lock all model inputs at a normal operating point, e.g. the power output is 90 MW, IGV position is 77% and etc., except the ambient variable being considered. For the simpler structure and closer variables relationship, the instance is chosen to use the single shaft unit, and consequently, Figure 9 illustrate the relative impact of these input parameters on the associated quality output measure for the CCGT plant.

It can be seen that increasing sea water temperature can significantly reduce the efficiency, the linear curve shows that about 50% increase in sea water temperature can cause 8% decrease in efficiency. Observably, the nonlinear curve shows that the relationship between

ambient conditions and efficiency is more complicated and non-monotonic. As shown in the red line in Figure 9, the effects of ambient air and sea water temperature on the plant efficiency are represented as two tendency directions. One is the temperature upon the 12 degree, the efficiency is decreased following the increase in the environment temperature, and the reason is well known as we discussed in section 6.1 that the reduced cooling water temperature can enhance the steam cycle efficiency. Another direction present an interesting result where the temperature is lower than 12 degree, the efficiency enhancement seems to be more difficult when the temperature goes down. The reason is considered that the air and fuels inlet temperature will be excessively decreased during chill period and causes the decrease in gas combustion temperature and consequently reduced the gas turbine efficiency; it tends to counteract the effects of decrease in cooling water temperature on efficiency enhancement.

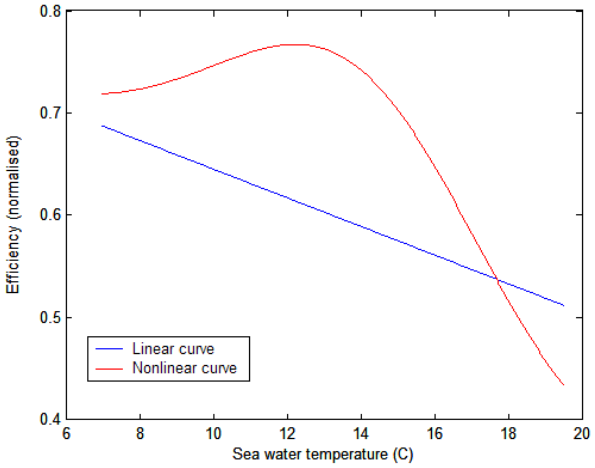


Fig. 9. Relation curve for single-shaft unit: efficiency vs. sea water temperature

7. Conclusion

Distributed control systems provide many advantages in terms of improvements in productivity and plant manoeuvrability when introduced into power plants and other industrial processes. However, the ease of access to a range of plant-wide signals potentially introduces vast problems of scale, since the meaningful information contained within the collected data may be somewhat less than the volume suggests. The task remains, therefore, to identify normal operating regions and relationships within the historical data, and subsequently to apply the collated rules, reference cases, etc. Principal component analysis has received considerable interest as a method of reducing the effective measurement space, and has been considered here for process monitoring of a combined cycle gas turbine.

Traditionally, operator practice has been reactive, whereby actions are taken following the triggering of process alarms, often set over-responsive and mode insensitive – PCA methods enable a more proactive role for the operator, providing early warning of plant irregularities, and identification of instrumentation errors and process faults. The PCA

model is identified under normal operating conditions, and subsequently unusual deviations are highlighted and identified.

On the other hand, The PLS model has received considerable interest as a method of analyzing process data and in this instance it has been used for analyzing a combined cycle gas turbine. Analysis of the models variance demonstrates that CCGT performance is affected by changes in ambient conditions. Also a relation curve method can be utilized here to study the impact of these external parameters, from the environment, have on the gas turbine performance.

Future work could extent the quality measurements (efficiency and emissions) to include other important requirements such as plant life, unit flexibility or cost of generation. Where the objective is to optimize power plant quality measurements and consequently enhance plant performance.

Furthermore, it is considered to integrate the physical or empirical mathematical models with the statistical model. Since the statistical models are limited by training range, it is unreliable to be employed under untrained conditions. Compared to physical model, a trained statistical model is difficult to tune in case of power plant renovation or update, due to it requests the input variables remaining a designed internal relationship.

On the other hand, developing the pure physical model can be a time-consuming exercise and requires the designer to have extensive knowledge of the application. Especially for the large scale system, the relationship between each variable is highly correlated and tangled, it is impossible to build a physical model which can provide a detailed and accurate representation of the operation. Therefore, utilizing the statistical model to substitute the complex components in a physical model could be good solution for reduce modeling calculation. Similarly, the adjustability and flexibility of statistical model can be increased by integrating some physical control loop or simulation into the statistical model.

8. References

Abdelhail, A., I. Traore and B. Sayed: 'RBDT-1: A New Rule-Based Decision Tree Generation Technique', *Rule Interchange and Applications*, 58, pp. 108-121, 2009

Ahvenlampi, T. and U. Kortela:' Clustering algorithms in process monitoring and control application to continuous digesters', *Informatica*, 29, pp.101-109, 2005

Armor, A. F.: 'Management and integration of power plant operations', *Thermal power plant simulation and control*, IEE Power and Energy Series 43, pp. 395-416, 2003

Baffi, G., E. B. Martin and A. J. Morris: 'Non-linear projection to latent structures revisited: the quadratic PLS algorithm', *Computers and Chemical Engineering*, 23, (3), pp. 395-411, 1999

Billings S.A and Hong, X.: "Dual-orthogonal radial basis function networks for nonlinear time series prediction". *Neural Networks*, 11 (3), pp. 479-493, April 1998

Blanco, C., D. Soronow and P. Stefiszyn: 'Multi-factor models for forward curve analysis: An introduction to principal component analysis', *Commodities Now*, June, pp. 76-78, 2002

Bharati M. H., J. J. Liu and J. F. MacGregor: 'Image texture analysis: methods and comparisons', *Chemometrics and Intelligent Laboratory Systems*, 72, PP. 57-71, 2004

Chan, C.: 'Application of multivariate analysis to Optimize Function of Cultured Hepatocytes', *Biotechnology Progress*, 19, (2), pp. 580-598, 2003

Cinar, A., and C. Undey: 'Statistical process and controller performance monitoring: A tutorial on current methods and future directions', *Proceedings of the American Control Conference*, San Diego, USA, 4, pp. 2625-2639, 1999

Dunia, R., S. Qin, T.F. Edgar and T.J. MCAVOY: 'Identification of faulty sensors using principal component analysis', *AIChE Journal*, 42, (10), pp. 2797-2812, 1996

Flynn, D., J. Ritchie, M. Cregan and L. Pan: 'Data mining techniques applied to power plant performance monitoring', *16th IFAC World Congress, Prague*, 2006.

Freund, R.J., and W.J. Wilson: 'Regression analysis: statistical modelling of a response variable', Academic Press Ltd., London, UK, 1997

Goldberg, E.D.: " Genetic Algorithms in Search, Optimization, and Machine Learning". *Addision-wesley publishing company*, INC. pp. 27-95, 1989

Jackson, D.A.: 'Stopping rules in principal component analysis: a comparison of statistical and heuristical approaches', *Ecology*, 74, (8), pp. 2204-2214, 1993

Jackson, J.E.: 'A user's guide to principal component analysis', Wiley series in probability and statistics, John Wiley & Sons, Hoboken, New Jersey, 2003

Jolliffe, I.T.: 'Principal component analysis', Springer series in statistics, New York, 2002

Hinkle, D. and C. Toomey: 'Applying case-based reasoning to manufacturing', *AI Magazine*, 16, (1), pp. 65-73, 1995

Kolodner, d.: 'Improving human decision making through case-based decision aiding', *AI Magazine*, 12, (2), pp. 52-68, 1991

Kourti, T., and J. F. Macgregor: 'Process analysis, monitoring and diagnosis, using multivariate projection methods', *Chemometrics and Intelligent Laboratory Systems*, 28, pp. 3-21, 1995

Kourti, T., J. Lee and J.F. Macgregor: 'Experiences with industrial applications of projection methods for multivariate statistical process control', *Computers and Chemical Engineering*, 20, pp. 745-750, 1996

Kresta, J.V., T.E. Martin and J.F. MacGregor: Development of inferential process models using PLS', *Computers and Chemical Engineering*, 18, (7), pp. 597-611, 1994

Lewin, D.R. 'Predictive maintenance using PCA', *Control Engineering Practice*, 3, (3), pp. 415-421, 1995

Li, W., H.H. Yue, S. Valle-Cervantes, and S.J. QIN: 'Recursive PCA for adaptive process monitoring', *Journal of Process Control*, 2000, 10, pp. 471-486

Li, G. and Liu, B.: "RBFNN algorithm based on hybrid hierarchy genetic algorithm and its application". *Control Theory & Applications.* 19(4), pp. 627-630, 2002.

MacGregor, J.F., and T. KOURTI: 'Statistical process control of multivariate processes' *Control Engineering Practice*, 3, (3), pp. 403-414, 1995

MacGregor, J.F., H. YU, S.G. Munoz and J. Flores-Cerrillo: 'Data-based latent variable methods for process analysis, monitoring and control', *Computers and Chemical Engineering*, 29, pp. 1217-1223, 2005

Martin, E.B., A.J. Morris and J. Zhang: 'Process performance monitoring using multivariate statistical process control', *IEEE Proceedings Control Theory and Applications*, 143, (2), pp. 132-144, 1996

Matthews, R.: 'Data miners only strike fool's gold', *New Scientist*, 8th March, pp. 8, 1997

Michalski, R.S., I. Bratko and M. Kubat" 'Machine learning and data mining: Methods and applications', Wiley & Sons, Chichester, England, 1999

Moody, Darken C.: "Fast Learning in Networks of Locally-tuned Processing Units". *Neural Computation*, 1, pp. 281 – 294,1989.

Olaru, C., and L. Wehenkel: 'Data Mining', *IEEE Computer Applications in Power*, 12, pp. 19-25, 1999

Oliveira-Esquerre K.P., D.E. Seborg, R.E. Bruns and M. Mori: 'Application of steady-state and dynamic modelling for the prediction of the BOD of an aerated lagoon at a pulp and paper mill Part I. Linear approaches', *Chemical Engineering Journal*, 104, pp. 73-81, 2004

Otto, M., and W. Wegscheider: 'Spectrophotometric multi-component analysis applied to trace metal determinations', *Analytical Chemistry*, 57, (1), pp. 63-69, 1985

Pan, L., D. Flynn and M. Cregan: 'Statistical model for power plant performance monitoring and analysis', *Universities Power Engineering Conference*, pp. 121-126, 2008

Qin, S., H. Yue, and R. Dunia: 'Self-validating inferential sensors with application to air emission monitoring', *Industrial Engineering Chemical Research*, 36, pp. 1675-1685, 1997

Quinlan, J.R.: 'C4.5: programs for machine learning', 1993, The Morgan Kaufmann series in machine learning, Morgan Kaufmann Publishers, California

Sebzalli, Y.M., and X.Z. Wang.: 'Knowledge discovery from process operational data using PCA and fuzzy clustering', *Engineering Applications of Artificial Intelligence*, 2001, 14, (5), pp. 607-616

Smyth, B., M.T. Keane and P. Cunningham: 'Hierarchical case-based reasoning integrating case-based and decompositional problem-solving techniques for plant-control software design', *IEEE Transactions on Knowledge and Data Engineering*, 13, (5), pp. 793-812, 2001

Valle, S., W. Li and S.J. Qin: 'Selection of the number of principal components: The variance of the reconstruction error criterion with a comparison to other methods', *Industrial Engineering Chemistry Research*, 38, pp. 4389-4401, 1999

Voumvoulakis, E.M. and Hatziargyriou, N.D.: 'A Particle Swarm Optimization Method for Power System Dynamic Security Control', *IEEE Transactions on Power Systems,* pp. 1032-1041, 2010

Watson, I.: 'Case-based reasoning is a methodology not a technology', *Knowledge-Based Systems*, 12, pp. 303-308, 1999

Wang, J.R.: 'Research on web-based multi-agent system for aeroengine fault diagnosis', *The 9th International Conference on Web-Age Information Management*, pp. 195-202, 2008

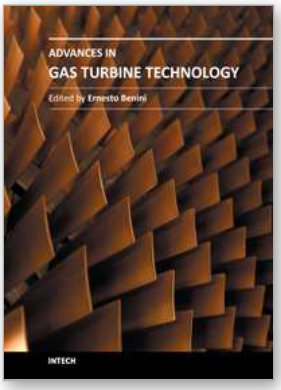
Weiss, S.M., and N. Indurkhya: 'Predictive data mining, a practical guide', 1998, Morgan Kaufmann Publishers Inc., San Francisco, CA

Wise, B.M. and N.B. Gallagher: 'The process chemometrics approach to process monitoring and fault detection', *Journal of Process Control*, 6, (6), pp. 329-348, 1996

Wold, S., K. Esbensen and P. Geladi: 'Principal component analysis', *Chemometrics and Intelligent Laboratory Systems*, 1987, 2, pp. 37-52

Yang, P.: 'A case-based reasoning with feature weights derived by BP network', *Workshop on Intelligent Information Technology Application, IITA*, pp. 26-29, 2007

Yoon, S., and J.F. Macgregor: 'Fault diagnosis with multivariate statistical models. Part one: using steady state fault signatures', *Journal of Process Control*, 11, pp. 387-400, 2001



Advances in Gas Turbine Technology

Edited by Dr. Ernesto Benini

ISBN 978-953-307-611-9
Hard cover, 526 pages
Publisher InTech
Published online 04, November, 2011
Published in print edition November, 2011

Gas turbine engines will still represent a key technology in the next 20-year energy scenarios, either in stand-alone applications or in combination with other power generation equipment. This book intends in fact to provide an updated picture as well as a perspective vision of some of the major improvements that characterize the gas turbine technology in different applications, from marine and aircraft propulsion to industrial and stationary power generation. Therefore, the target audience for it involves design, analyst, materials and maintenance engineers. Also manufacturers, researchers and scientists will benefit from the timely and accurate information provided in this volume. The book is organized into five main sections including 21 chapters overall: (I) Aero and Marine Gas Turbines, (II) Gas Turbine Systems, (III) Heat Transfer, (IV) Combustion and (V) Materials and Fabrication.

How to reference

In order to correctly reference this scholarly work, feel free to copy and paste the following:

Li Pan (2011). Application of Statistical Methods for Gas Turbine Plant Operation Monitoring, Advances in Gas Turbine Technology, Dr. Ernesto Benini (Ed.), ISBN: 978-953-307-611-9, InTech, Available from: http://www.intechopen.com/books/advances-in-gas-turbine-technology/application-of-statistical-methods-for-gas-turbine-plant-operation-monitoring

INTECH
open science | open minds

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821

© 2011 The Author(s). Licensee IntechOpen. This is an open access article distributed under the terms of the Creative Commons Attribution 3.0 License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.